Anaforafeloldás magyar nyelvű szövegekben

Lejtovicz Katalin Eszter[1], Kardkovács Zsolt Tivadar[1]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék,
H-1117 Budapest, Magyar Tudósok krt. 2.

1 Bővített kivonat

Napjainkban az anaforafeloldás problémájának algoritmikus megoldása egyre inkább döntő fontosságú, és sürgető feladat, mivel számos alkalmazásban szükség van az utalószavak gyorsan és helyesen történő feloldására. Gépi fordítás esetén fontos, hogy a személyes névmási anaforákat - amennyiben a célnyelv megkülönböztet nemeket - a célnyelv megfelelő nemű személyes névmási anaforájára fordítsuk. Tény-kinyerést megvalósító algoritmusok implementálásakor figyelembe kell venni, hogy hosszabb szövegek esetén a szöveg elején elhangzott központi témára később sokszor utalószóval hivatkozunk. A keresők, illetve indexelők esetében a találati pontosság 15-20%-os hibájának egyik legmeghatározóbb oka, hogy a szövegben található anaforikus elemek különböző kifejezésekre képződnek le, ezért a szógyakoriságon alapuló indexelés eredő hibája és bizonytalansági tényezője relatív magas. Általánosságban tehát azt mondhatjuk, hogy a helyes anaforafeloldás bármilyen szövegfeldolgozással kapcsolatos területen kulcskérdés.

A feloldásra használt algoritmus kiválasztása esetén figyelembe kell venni, hogy milyen fajta anaforák fordulnak elő a magyar nyelvben, és hogy az egyes típusok milyen gyakorisággal jelennek meg a szövegekben. Az anaforák hat legfontosabb típusa a következő: névmási, zéró, határozói, NP, igei, rész-egész (lásd 1. táblázat). A felsorolás egyben tükrözi az egyes típusok előfordulási gyakoriságát is, balról jobbra haladva az egyre csökkenő gyakoriságú anaforákat tüntettük fel.

Az anaforafeloldási algoritmusok két fő típusba sorolhatóak, azaz megkülönböztetünk tudás alapú és tudásszegény algoritmusokat. A tudás alapú rendszerek emberi munkával előfeldolgozott bemeneten dolgoznak, a tudásszegény módszerek azonban automata parszolást végeznek. A tudás alapú feldolgozás jóval megbízhatóbb eredményeket ad mint a tudásszegény, viszont az emberi munka felhasználásából következően lassabb és költségesebb is annál. Ezidáig anaforafeloldást megvalósító algoritmusok főként angol nyelvre születtek. Magyar nyelvre a poszterben is bemutatásra kerülő program használható. Ez az algoritmus a tudás alapú kategóriába tartozik, és ezen belül is az úgynevezett CT (Centering Theory) elvet használja.

A CT a következőképpen foglalható össze:

1 Egy diskurzusban mindenegyes megnyilatkozásnak pontosan egy központi témája van.
2 Egy anafora nagy valószínűséggel a központi témára utal vissza.

A diskurzus során az egymást követő megnyilatkozások általában az előző mondat központi témáját folytatják.

~ anafora	Magyar nyelvű példa mondatok
névmási	*Péter* azért nem jött el a moziba, mert **ő** már látta a filmet.
zéró	A *lány* elment a boltba, de **(ő)** nem vitt magával pénzt.
határozói	Mi elmegyünk a *vendéglőbe*, és veled majd **ott** találkozunk.
NP	*Bush* felszólalt a szenátusban. **Az elnök** beszédében…
igei	A *kislány* énekelt, és a testvére is **így tett**.
rész-egész	Bár *Svájcban* 1,2€ a benzin litere, **Zürich**ben jóval drágább.

1. táblázat. Az anaforák típusai.

Az egyik legjobban ismert CT alapú algoritmus, Brenan, Friedman és Pollard 1987-es (röviden BFP) algoritmusa is a középpontba helyezés elvét használja.

A szakirodalomból ismert algoritmusok közül azért a BFP-t érdemes a magyar nyelvre alkalmazhatóvá tenni, mivel ez az algoritmus figyelembe veszi nyelvünk sajátosságait. A BFP algoritmus előnye, hogy jó találati aránnyal működik mind az izoláló típusú angol nyelvre, mind az agglutináló magyarra. Tehát jól működik a mondatrészek sorrendjét kevésbé megkötő magyar mondatok esetében, és a sorrendet jobban megkötő angol mondatok esetében is. A BFP algoritmus magyar nyelvre történő adaptálásával a leggyakrabban előforduló anaforák, vagyis a névmási, zéró és határozói anaforák feloldása valósítható meg. A program bemeneteként a Szeged Korpusz CD-n található nyelvtanilag elemzett mondatok szolgáltak. A CD-n található magyar nyelvű szövegek szintaktikailag és morfológiailag is elemzettek. Az algoritmus anafora-antecedens párok képzését, a biztosan rossz jelöltek kiszűrését és a megmaradtak közül a mondatok közötti átmenetek rangsorolása után kapott, legnagyobb valószínűséggel helyes pár kiválasztását végzi. A programot leginkább már csak nyelvészeti területen kell fejleszteni (szűrésben szereplő feltételek szigorításával, új információ-régi információ szerinti tagolással). A program tesztelése a Szeged Treebank 2.0 CD-n történt.

A kapott eredmények:

- A szövegben levő összes anaforának a 37%-át találja meg a program.
- A szövegben levő azon anaforákból, melyek megtalálását a programban megkíséreljük, 39,6% megtalálása sikeres
- A program 21%-ot old fel helyesen az összes anaforából
- A program 23,8%-ot old fel helyesen azokból az anaforákból, amelyek feloldását megkíséreljük

A magyarra megvalósított BFP algoritmus találati arányát (39,6%) összehasonlítva az angolra készült BFP algoritmuséval (59%) azt állapíthatjuk meg, hogy az angol nyelv esetében kb. 20%-kal értek el jobb eredményeket. Nagy valószínűséggel a már említett továbbfejlesztések hatására a magyarra írt program feloldási aránya el fogja érni az angolnál tapasztaltakét.